UNIVERSITAT DE BARCELONA

Grau de Lingüística

Treball de Fi de Grau

Curs 2021-2022

TÍTOL: Detecció automàtica de la ironia en espanyol

NOM DE L'ESTUDIANT: Emma Raimundo Schulz

NOM DEL TUTOR: Mariona Taulé

Barcelona, setembre 2022

Agraïments

Vull donar les gràcies a la Mariona Taulé, la meva tutora, per guiar-me durant el procés de desenvolupament d'aquest treball, i oferir-me suport, consell i acompanyament davant els moments de dificultat. Agraeixo de tot cor la seva confiança en mi i el seu interès per aquest projecte.

A l'Alejandro Ariza, per haver dedicat el seu temps a introduir-me al món de l'aprenentatge automàtic i haver-me motivat a seguir formant-me en el llenguatge de programació Python i a continuar explorant els àmbits del processament del llenguatge natural i la intel·ligència artificial.

Finalment, a la meva família i amics, per escoltar-me pacientment i estar pendents del progrés d'aquest projecte. Especialment, al meu pare, que sense voler-ho ha resolt amb els seus consells algunes de les complicacions que han sorgit durant la implementació dels models.

Resum

La detecció automàtica del llenguatge figurat està generant cada vegada més interès en l'àmbit de la lingüística computacional. Aquest estudi, emmarcat dins els àmbits del Processament del Llenguatge Natural (PLN) i l'Aprenentatge Automàtic, documenta el procés de construcció d'un classificador de la ironia en textos en espanyol a partir de models com un Support Vector Machine (SVM) i BERT multilingüe. Els millors resultats obtinguts pertanyen al SVM, amb un valor F_1 de 0.85 punts. Pel que fa al marc teòric, es parteix de les definicions d'ironia i sarcasme que ofereixen autores com Reyes (1994) i Ruiz Gurillo (2012) i de les consideracions de Kerbrat-Orecchioni (1981) i Reus Boyd-Swan (2009), que suggereixen la presència de diverses marques lingüístiques en el to irònic.

Paraules clau: ironia, sarcasme, Aprenentatge Automàtic, Processament del Llenguatge Natural.

Abstract

The automatic detection of figurative language is increasingly generating more interest in the field of computational linguistics. This study, framed within the fields of Natural Language Processing and Machine Learning, documents the building process of an automatic irony classifier in Spanish texts based on models such as a Support Vector Machine (SVM) and multilingual BERT. The best results are obtained with the SVM, with an F1 value of 0.85 points. Regarding the theoretical framework, we consider the definitions of irony and sarcasm offered by authors such as Reyes (1994) and Ruiz Gurillo (2012) and the ideas of Kerbrat-Orecchioni (1981) and Reus Boyd-Swan (2009), which suggest the presence of several linguistic marks in the ironic tone.

Keywords: irony, sarcasm, Machine Learning, Natural Language Processing.

UNIVERSITAT DE BARCELONA

Coordinació d'Estudis
Facultat de Filologia i Comunicació

Gran Via
de les Corts Catalanes, 585
08007 Barcelona

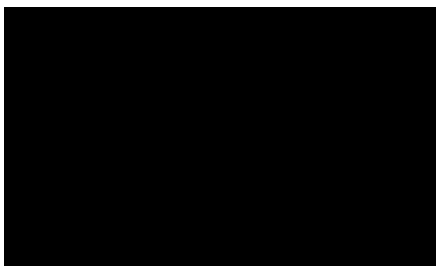
Tel. +34 934 035 594
fil-coord@ub.edu
www.ub.edu

Declaració d'autoria

Amb aquest escrit declaro que soc l'autor/autora original d'aquest treball i que no he emprat per a la seva elaboració cap altra font, incloses fonts d'Internet i altres mitjans electrònics, a part de les indicades. En el treball he assenyalat com a tals totes les citacions, literals o de contingut, que procedeixen d'altres obres. Tinc coneixement que d'altra manera, i segons el que s'indica a l'article 18 del capítol 5 de les Normes reguladores de l'avaluació i de la qualificació dels aprenentatges de la UB, l'avaluació comporta la qualificació de "Suspens".

Barcelona, a 5 d'agost de 2022

Signatura:



Emma Raimundo Schulz

ÍNDEX

1. INTRODUCCIÓ

A les darreres dècades, el desenvolupament de les Tecnologies de la Informació i la Comunicació ha generat l'aparició de diversos gèneres específics del discurs electrònic, pertanyents a la Web 2.0: comentaris digitals, hipertextos, blogs, textos en pàgines web, posts en xarxes socials i fins i tot mems. Aquests gèneres discursius, alguns possiblement híbrids entre verbals i escrits, pertanyen al tipus de Comunicació Mitjançada per Ordinador (Herring, 1996) i tenen com a mitjà el ciberespai o mitjà electrònic (anomenat "tercer mitjà" per Crystal (2011)).

Des de la perspectiva de la lingüística computacional i el Processament del Llenguatge Natural (PLN), aquesta explosió informativa planteja la necessitat de solucions eficients i fiables per poder gestionar, analitzar i extreure coneixement de grans quantitats de dades (Potamias et al., 2020). Pel que fa a la lingüística computacional, cada vegada hi ha més interès en el processament del llenguatge figurat, i una de les tasques que està generant interès en els últims anys és la detecció automàtica de la ironia, ja que encara no hi ha una solució general en un únic algorisme o tècnica que detecti aquest fenomen lingüístic en els textos escrits (Zhang i Abdul-Mageed, 2019).

La ironia es fa servir sovint per expressar opinions. És per això que detectar-la forma part de la preocupació dels investigadors de l'àrea d'anàlisi de xarxes socials i de l'anàlisi de sentiments, que fan servir la detecció de la ironia per evitar la confusió dels tons irònics com a missatges literals (Calvo et al., 2020). A més, també pot ser interessant per a la detecció del discurs d'odi (Taulé et al., 2021) i la detecció de notícies falses (*fake news*), entre altres (Zhang i Abdul-Mageed, 2019).

Tradicionalment, des de la pragmàtica es defineix la ironia com la infracció de la màxima de la qualitat, perquè sembla dir justament el contrari del que es vol expressar. Tot i això, Reyes (1994) defineix la ironia com un cas especial del tipus de cita que alguns autors anomenen ressò (*eco*). L'autora apunta que els ressons irònics són una repetició d'un enunciat previ (o del contingut) i que es poden fer servir per assenyalar conformitat amb aquest enunciat o per indicar que s'ha entès. En la ironia, el parlant es fa ressò del contingut d'un altre enunciat deformant-lo, exagerant-lo o modificant-lo amb to de burla per mostrar una actitud negativa davant d'aquest enunciat o cap al seu autor (p.ex. "Quin dia tan meravellós!" Durant un dia de

pluja). El sarcasme, en canvi, constitueix la intensificació d'aquesta crítica (p. ex. “La veritat que ets un geni…”).

Segons Reyes (1994), la funció discursiva dels ecos irònics és mostrar incongruència entre la proposició esmentada pel parlant i la situació present. Aquest contrast causa un desajust que resulta xocant o inadequat, provocant moltes vegades efectes humorístics. A més, la comprensió de la ironia es converteix en un joc plaent, on l'interlocutor obliga el receptor a construir en comú uns significats. Aquest procés posa en relleu la complicitat entre tots dos i reforça la relació entre ells.

Per la seva part, Ruiz Gurillo (2012) destaca que on hi ha ironia hi ha desdoblament del locutor. Segons l'autora, la ironia constitueix una representació en la qual el llest repeteix les paraules del ximple, però allunyant-se d'elles i mostrant la seva actitud davant aquelles paraules i davant la situació en la qual es troba. Per aquesta raó, podem parlar de polifonia, d'una coexistència de veus en un mateix enunciat, ja que l'autor de la ironia no es responsabilitza de la veritat literal de la seva proposició, sinó que la responsabilitat del missatge s’assigna a un alter ego ridícul.

Tot i que el desajust entre el contingut de l'expressió i la situació real és allò que permet entendre el to irònic o sarcàstic, l'autor d'una ironia o sarcasme aconsegueix manifestar-la també per mitjà del to de veu, gestos, o el registre utilitzat (Reyes, 1994). No obstant això, en el context del mitjà digital, la comprensió de la ironia en textos escrits no recau en les marques cinèsiques, les marques paralingüístiques i de caràcter acústic-melòdic, tal com assenyala Reus Boyd-Swan (2009). Considerem en aquest treball que en el medi cibernètic la falta d’informació inferida a través de la prosòdia o altres elements paralingüístics és compensada pels autors a través de marques lingüístiques com emoticones, *hashtags* i signes de puntuació.

Emmarquem aquest estudi dins l’àrea del PLN, aplicant una metodologia basada en l’aprenentatge automàtic (Machine Learning, ML) per construir un classificador automàtic d’ironia en espanyol amb els models Support Vector Machine (SVM) i BERT. A més, partim de les definicions d’ironia i sarcasme que ofereixen autores com Reyes (1994) i Ruiz Gurillo (2012) i de les consideracions de Kerbrat-Orecchioni (1981) i Reus Boyd-Swan (2009), que suggereixen la presència de diverses marques lingüístiques en el to irònic.

L'estructura d'aquest treball és la següent: A l'apartat 2 es recopilen diversos antecedents en la detecció automàtica de la ironia. L'apartat 3 presenta els objectius i la hipòtesi del treball. A l'apartat 4 es descriu la metodologia emprada, es formula la tasca i s'explica detalladament tot el procés d'aprenentatge automàtic (selecció de les dades, preprocessament i vectorització, implementació dels models i l'avaluació d'aquests). A l'apartat 5 es comparen els resultats obtinguts i es proposen altres mètodes o tècniques per a la resolució d'aquesta tasca. Finalment, el treball conclou a l'apartat 6.

2. ANTECEDENTS

Aquest apartat recopila diferents estudis que tracten la detecció automàtica de la ironia en les últimes dècades. Igual que en altres tasques de PLN, la major part de la recerca sobre aquest fenomen lingüístic s'ha realitzat en textos en anglès, provinents de plataformes en línia com Amazon, fòrums, Reddit i Twitter (Ortega-Bueno et al., 2019).

La tasca de detecció de la ironia, considerada un problema de detecció binària, s'ha abordat de diferents maneres al llarg dels anys. En estudis com els de Carvalho et al. (2009), González-Ibáñez et al. (2011) i Kunneman et al. (2015) s'han fet servir *features* (trets) basats en el text, com n-grames i etiquetes morfosintàctiques resultants de la desambiguació lèxica (*POS tagging*), per representar i identificar missatges irònics. No obstant això, la ironia sovint no es manifesta explícitament amb marques lèxiques i està vinculada a aspectes subjectius. Per aquesta raó, estudis com Farías et al. (2016) desenvolupen un model de detecció automàtica de la ironia en tuits en anglès a partir de l'anotació d'informació afectiva de cada text (fent servir recursos com el Dictionary of Affect in Language, i lexicons com Hu&Lui i AFINN), a més de prestar atenció a l'estructura d'aquests (signes de puntuació, llargària del text i emoticones entre altres).

També existeixen estudis que han aprofitat recursos d'anàlisi de sentiments per analitzar la presència d'ironia en les xarxes, com Irazú et al. (2015), que duen a terme una tasca d'anàlisi de sentiments en Twitter, a partir d'un model de regressió que assigna valors de polaritat a cada tuit, que indiquen un nivell major o menor d'ironia. Finalment, cal mencionar també les aproximacions basades en l'aprenentatge profund (*Deep Learning*), ja que les incrustacions de paraules (*word-embeddings*) i les xarxes neuronals convolucionals s'han explotat per captar la presència d'ironia en els textos de les xarxes socials Ortega-Bueno et al. (2019). Per

exemple, en Potamias et al. (2020), es proposa una metodologia per a la detecció de la ironia en anglès basada en la implementació d'una xarxa neuronal preentrenada.

Pel que fa a l'espanyol, la baixa disponibilitat de corpus anotats amb ironia ha limitat la quantitat de recerca en aquest àmbit (Ortega-Bueno et al., 2019). Es presenten a continuació en ordre cronològic alguns dels estudis més rellevants sobre la detecció de la ironia en l'espanyol en l'última dècada.

Barbieri et al. (2015) estudien la caracterització dels missatges de Twitter en castellà que anuncien notícies satíriques i implementen un model SVM per a la detecció de la ironia basant-se en la selecció de *features* com la freqüència de paraules, l'ambigüitat de les paraules, la freqüència de categories morfosintàctiques, sinònims, sentiments, caràcters i paraules d'argot. Similarment, Jasso i Meza (2016) implementen els algorismes SVM i Random Forest per un model de classificació basat en *features* de paraules i caràcters per a detectar la ironia en tuits escrits en castellà.

Ortega-Bueno et al. (2019) introdueixen la primera tasca de col·laboració dedicada a detectar la ironia en tuits i comentaris a notícies en diferents variants de l'espanyol en el marc del Workshop IberLEF 2019: l'IroSvA[1]. Aquesta tasca estudia la manera en la qual la ironia canvia entre les diferents variants de l'espanyol que es parla a Espanya, Mèxic i Cuba. Calvo et al. (2020), a partir de la data proporcionada per l'IroSvA, proposen un model simple per la identificació de la ironia en tres variants de la llengua espanyola, basat en *embeddings* de tuits. Per a la generació d'aquests *embeddings*, tenen en compte cinc graus d'intensitat (cap, molt baixa, baixa, alta i molt alta) respecte de sis *features* (amor, alegria, sorpresa, tristesa, ira i por) en funció del grau d'emoció present en cada tuit.

Des d'una perspectiva multilingüe, Cignarella et al. (2020), exploren una varietat de funcions basades en dependències sintàctiques combinades amb classificadors clàssics d'aprenentatge automàtic per a la detecció de la ironia en espanyol, anglès, francès i italià, a més d'implementar el model d'aprenentatge automàtic BERT multilingüe basat en xarxes neuronals (cita referència dels creadors de BERT multilingüe).

Més recentment, Alnajjar & Hämäläinen (2021) han generat el primer corpus de sarcasme (un tipus d'ironia) multimodal, que combina vídeo i àudio, en espanyol peninsular i espanyol

[1] https://www.autoritas.net/IroSvA2019/

llatinoamericà. Presenten també diversos models per a la detecció del sarcasme, entre ells un SVM.

3. OBJECTIUS I HIPÒTESI

L'objectiu general d'aquest treball és, d'una banda, aprendre com és el procés de desenvolupament d'un classificador que resol una tasca lingüística i que requereix l'aplicació de tècniques d'Aprenentatge Automàtic (Machine Learning) i eines de PLN. El propòsit és conèixer i aplicar la metodologia necessària per a la construcció del classificador, des de la generació de dades, gestió del corpus i processament de les dades fins a la implementació i avaluació dels diferents models supervisats d'aprenentatge automàtic utilitzats. Aquest treball permetrà aplicar, ampliar i aprofundir els coneixements adquirits durant el grau de Lingüística sobre el llenguatge de programació Python i diferents eines de PLN.

D'altra banda, l'objectiu concret d'aquest treball és la construcció d'un classificador que sigui capaç de detectar la ironia en comentaris digitals i tuits en espanyol. La hipòtesi de partida és la següent: és possible classificar automàticament els textos en espanyol segons si contenen una ironia o no, aplicant mètodes d'aprenentatge automàtic supervisat, és a dir, models entrenats a partir de corpus textuals prèviament anotats.

4. METODOLOGIA

Aquest estudi documenta el procés de construcció d'un classificador de la ironia en textos en espanyol extrets de tuits i de comentaris a notícies. Per portar a terme aquesta tasca, ens basem en la metodologia proposada per Hladka i Holub (2015), que consisteix en la formulació de tasques, el desenvolupament de l'aprenentatge automàtic i l'obtenció del classificador. Durant el procés d'aprenentatge automàtic, s'obtenen les dades, es construeix el classificador i s'avalua el model implementat. El classificador de la ironia que es presenta en aquest treball s'ha construït a partir de diverses llibreries de Python dissenyades per a portar a terme tasques de PLN i aprenentatge automàtic, com Pandas, Numpy, Sklearn i Tensorflow[2].

[2] Els scripts generats per al desenvolupament d'aquest treball s'han adjuntat als annexos.

En la primera fase de desenvolupament del projecte, les dades que s'han fet servir pertanyen al corpus NewsCom-TOX (Taulé et al., 2021), amb les quals s'ha entrenat un model supervisat anomenat Support Vector Machine (SVM). Tal com s'especificarà més endavant, els resultats poc satisfactoris que s'han aconseguit a partir de les dades obtingudes del corpus NewsCom-TOX han portat a la selecció d'un altre corpus dissenyat específicament per aquesta tasca, és a dir, per a la detecció de la ironia, i amb dades més equilibrades. En aquesta segona fase, s'han utilitzat dades recuperades a partir del corpus d'ironia de Jasso i Meza (2016), que s'han preprocessat per entrenar dos models d'aprenentatge automàtic: un SVM i el transformador BERT.

Aquest apartat comença formulant breument la tasca que s'ha dut a terme. Les seccions a continuació estan estructurades segons els diferents tipus de dades utilitzades. En ambdós casos, es descriuen les especificacions dels corpus, les tècniques i criteris de preprocessament i processament de les dades que requereixen els models implementats (vectorització) i els paràmetres d'implementació d'aquests. Finalment, es mostren els resultats obtinguts per cada corpus aplicant mètodes d'avaluació com matrius de confusió i els valors d'F_1.

4.1 Formulació de la tasca

La tasca que s'ha dut a terme en aquest treball és la construcció d'un classificador que detecta automàticament la ironia en textos digitals en espanyol. Aquest classificador resoldrà una tasca de classificació binària: serà capaç d'assignar un valor a l'input (el missatge, p.ex. un tuit) segons si conté ironia o no. Cada text que rep serà classificat com no irònic (valor 0) o irònic (valor 1).

Cal destacar que en la utilització del corpus NewCom-TOX, la detecció del classificador es limitarà al to sarcàstic, definit com un tipus d'ironia. Amb aquest corpus, el SVM implementat s'entrenarà amb comentaris digitals provinents de notícies web. En canvi, amb el corpus de Jasso i Meza (2016), anomenat corpus IRONIA en aquest treball, conformat per tuits en espanyol, el classificador identificarà el to irònic. Es mostren a continuació exemples de textos pertanyents a ambdós corpus (Figura 1), cadascun anotat segons dels seus propis criteris, que seran detallats en els subapartats següents. La detecció d'ironia del classificador, per tant, variarà en funció de la definició que adopta cada corpus.

Corpus	Etiqueta	Text
NewsCom-Tox (2021)	Sarcasme	*Son sus costumbres y hay que respetarlas...*
IRONIA (2016)	Ironia	*Bien lucia bien, cada día mas lista* 👏👏 *#ironia*

Figura 1. Comparació d'instàncies dels corpus NewsCom-TOX (2021) i IRONIA (2016)

4.2 Corpus NewsCom-TOX

Segons Hladka & Holub (2015), per obtenir un *dataset* adequat cal recopilar dades amb una classificació assignada, netejar-les i preprocessar-les. Les dades utilitzades inicialment per a la construcció del classificador provenen del corpus NewsCom-TOX, que recull aproximadament 4,353 comentaris digitals en espanyol peninsular d'articles publicats en diaris en línia espanyols (ABC, elDiario.es, El Mundo, NIUS, etc.) i fòrums de discussió (com Menéame) entre els anys 2017 i 2020 (Taulé et al., 2021). Els textos d'aquest corpus estan seleccionats a partir d'un mètode de cerca de paraules clau (*keyword approach*) relacionades principalment amb articles d'immigració, i estan anotats manualment per tres anotadors segons paràmetres de toxicitat. La versió final del corpus conté l'acord entre els anotadors, també anomenat *Gold Standard* (DETOXIS-IberLEF, 2021)[3].

El paràmetre seleccionat per a la tasca d'aquest treball és l'etiqueta "*sarcasm*". El sarcasme es defineix en aquest corpus a partir de la definició tradicional de la ironia: ocorre quan el parlant diu el contrari a allò que vol expressar, acompanyat d'una crítica dura i negativa. A més, té anotades com a sarcàstiques les preguntes retòriques, els jocs de paraules i les bromes acompanyades de crítica o burla (Taulé et al., 2021). Dels més de quatre mil comentaris digitals que conformen aquest corpus, 460 estan anotats com sarcàstics (10,6%) i 3893 com a no sarcàstics (89,4%).

4.2.1 Selecció de dades i preprocessament

Les dades del corpus NewsCom-TOX (Taulé et al., 2021) van ser proporcionades pel grup de recerca Centre de Llenguatge i Computació (CLiC)[4] de la Universitat de Barcelona. S'ha tingut accés als arxius preprocessats i separats en *training* i *test*, ja que aquestes dades s'han

[3] https://detoxisiberlef.wixsite.com/website
[4] http://clic.ub.edu/ca/

compartit prèviament per desenvolupar models d'aprenentatge automàtic que detectin la toxicitat[5].

Separar les dades entre *training* i *test* permet avaluar el rendiment d'algorismes d'aprenentatge automàtic. Aquesta tècnica consisteix a dividir el corpus, la base de dades, en dos subconjunts: el primer es fa servir per entrenar el model (*training dataset*) i el segon per avaluar-lo (*testing dataset*). Una vegada el model està entrenat amb les dades d'entrenament, es proveeix al model amb les dades d'avaluació i l'algorisme prediu els valors que se li demanen. Els valors predits, el seu output, es comparen amb els valors vertaders (és a dir, els anotats i consensuats manualment pels anotadors, el corpus *gold standard*). Per tant, la separació entre *train* i *test* de les dades permet avaluar el rendiment de l'algorisme (Brownlee, 2020).

Entre els diferents paràmetres de toxicitat anotats en el corpus, s'han seleccionat per aquesta tasca la categoria "*sarcasm*" exclusivament. S'observa en la Figura 2 exemples de comentaris pertanyents al corpus NewsCom-TOX.

Comentari	Sarcasme
"Les pagaran ellos de su bolsillo ,si no se llenara España de inmigrantes ilegales total estos los legalizaran"	0
"También podemos hacerlo con todos los venezolanos que han tenido que huir de su país, de su casa, por la violencia chavista... pero, ah, no esos no que no les votarían a ellos, son más de delincuente común."	1

Figura 2. Exemples de textos del corpus NewsCom-TOX (Taulé et al., 2021).

4.2.2 Vectorització de les dades: TF/IDF

Els algorismes d'aprenentatge automàtic requereixen un input numèric, com una matriu bidimensional amb files com a instàncies i columnes com a característiques. Per tant, en una tasca d'aprenentatge automàtic que implica dades en llenguatge natural és necessari transformar les dades de format text a un format que el model pugui interpretar. Aquest procés, que converteix el text en vectors numèrics, s'anomena vectorització (Ramadhan, 2021).

[5] NewsCom-TOX es va usar com a corpus d'aprenentatge i avaluació en la tasca DETOXIS en el marc del Workshop IberLEF-2021: https://detoxisiberlef.wixsite.com/website

Per a dur a terme la vectorització de les dades en aquest estudi s'ha implementat, des de la llibreria Sklearn de Python, el TF-IDF Vectorizer, que combina 2 conceptes: *Term Frequency* (TF) i *Document Frequency* (DF). El valor de *Term Frequency* fa referència al nombre d'ocurrències d'un terme específic en un text del corpus. Aquest valor indica la importància de cada paraula dins del text. El *Term Frequency* representa cada text de les dades com una matriu formada per files que representen cada text i columnes que representen tots els termes (tokens) que apareixen al corpus (Figura 3) (Ramadhan, 2021).

	and	but	hate	i	image	language	like	love	natural	processing	python	signal
Text 1	0	1	1	2	0	1	0	1	1	1	1	0
Text 2	0	0	0	1	1	0	1	0	0	1	0	0
Text 3	1	0	0	1	1	0	1	0	0	2	0	1

Figura 3. Exemple Term Frequency (Ramadhan, 2021).

En canvi, el *Document Frequency* és la quantitat de documents del corpus que contenen una paraula específica. Aquest valor indica com és de comú un terme del corpus. A partir dels valors del DF, es calcula el *Inverse Document Frequency* (IDF), que mostra el pes de les paraules del corpus. S'assigna un pes més alt als termes rars o poc freqüents i un pes més baix als termes més comuns i més escampats pels documents del corpus (Figura 4) (Ramadhan, 2021).

Term	and	but	hate	i	image	language	like	love	natural	processing	python	signal
IDF	0.47712	0.47712	0.4771	0	0.1760913	0.477121	0.1760913	0.477121	0.47712125	0	0.477121	0.477121

Figura 4. Exemple Inverse Document Frequency (Ramadhan, 2021).

Finalment, es multipliquen els valors de TF i IDF, que resulta en una matriu amb valors numèrics (Figura 5).

	and	but	hate	i	image	language	like	love	natural	processing	python	signal
Text 1	0	0.47712	0.4771	0	0	0.477121	0	0.477121	0.47712125	0	0.477121	0
Text 2	0	0	0	0	0.1760913	0	0.1760913	0	0	0	0	0
Text 3	0.47712	0	0	0	0.1760913	0	0.1760913	0	0	0	0	0.477121

Figura 5. Exemple TF-IDF (Ramadhan, 2021).

4.2.3 Support Vector Machine

El model d'aprenentatge automàtic SVM constitueix una aproximació geomètrica a les tasques d'aprenentatge automàtic. Aquest mètode distribueix les dades (*features*) en l'espai i

construeix un hiperplà òptim, un límit de decisió lineal, que separi les instàncies en dues classes. Matemàticament, un hiperplà pren la forma d'una funció lineal $f(x)=wx + b$, on els paràmetres w i b representen paràmetres d'hipòtesi del classificador SVM. Un cop entrenat el model, es troba l'hiperplà que separa les diferents classes. A continuació, les noves instàncies que es proveeixin al model SVM poden ser classificades segons la seva posició en l'espai (Hladka i Holub, 2015). En la Figura 6, les instàncies situades per sobre de l'hiperplà seran classificades com estrelles i les situades per sota com triangles.

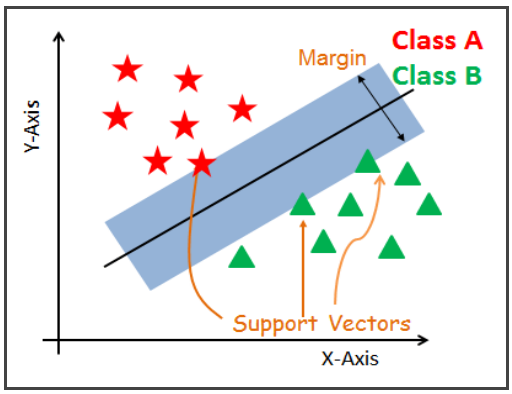


Figura 6. Representació gràfica en dues dimensions de la distribució espaial dels features en un SVM i l'hiperplà que les separa (Navlani, 2019).

Segons (Hladka i Holub, 2015), poden existir una infinitat d'hiperplans de separació entre els quals l'algorisme SVM selecciona el que sigui més distant a les instàncies més properes de diferents classes. Aquest hiperplà es pot anomenar *maximum marginal hyperplane* (MMH) (Navlani, 2019). Malgrat que el SVM s'implementa majoritàriament per tasques de classificació binària, una tasca multiclasse es pot reduir a diverses tasques binàries "un contra tots" (*one-versus-all*) o "un a un" (*one-to-one*) (Hladka i Holub, 2015).

Implementació del model

L'algorisme de SVM s'implementa fent servir un kernel, que transforma les dades d'entrada (la matriu de vectors) a la forma requerida (Navlani, 2019). A més, els diferents paràmetres de kernel seleccionen el tipus d'hiperplà que separa les dades (Ben Fraj, 2018).

En aquest estudi, hem implementat el SVM amb un kernel lineal i amb un kernel RBF, a partir de la llibreria Sklearn de Python (Annex 1). El kernel lineal (Figura 9) pot resoldre problemes de classificació binària i separa classes de dades linealment separables (Figura 7). En canvi, si les dades d'entrenament no són linealment separables (Figura 8), es poden fer

servir les funcions de kernel que converteixen les dades en separables distribuint els *features* en l'espai en una dimensió superior, en la qual poden aparèixer propietats no lineals (Hladka & Holub, 2015).

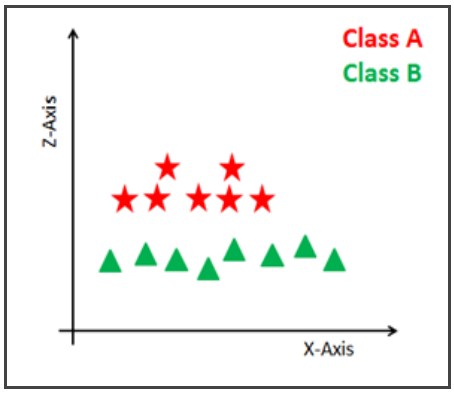


Figura 7. Features linealment separables (Navlani, 2019).

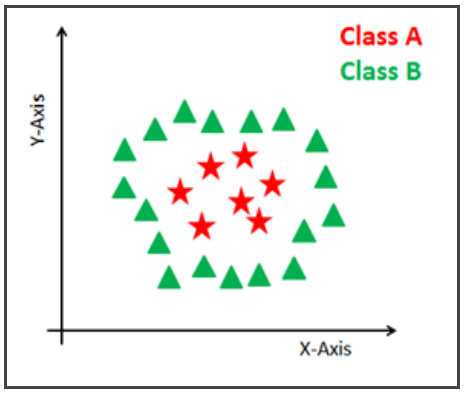


Figura 8. Features no separables linealment (Navlani, 2019)

El kernel RBF (Radial Basis Function Kernel) es fa servir sovint en tasques de classificació, ja que permet distribuir les dades en l'espai en infinites dimensions (Navlani, 2019) i trobar hiperplans més complexos quan les dades no es poden separar linealment (Figura 10).

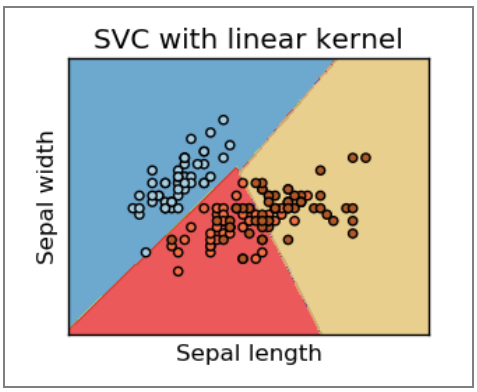


Figura 9. Hiperplà del SVM amb Kernel Lineal (Ben Fraj, 2018).

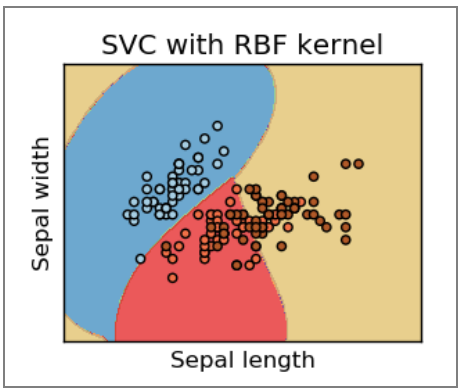


Figura 10. Hiperplà del SVM amb Kernel RBF (Ben Fraj, 2018).

4.2.4 Avaluació del model

Per a l'avaluació final del classificador, se li ha entregat com a *input* al SVM la *testing data*, és a dir, els textos o instàncies que són desconegudes per l'algorisme i que no contenen la classificació real (del *gold standard*). Les prediccions que ofereix el SVM, l'*output*, han estat avaluades amb una matriu de confusió i el valor F_1. Les matrius de confusió són el mètode

principal per avaluar problemes de classificació, ja que informen sobre les instàncies que l'algorisme ha classificat errròniament. En una classificació binària (0-1, com en el nostre cas), la matriu de confusió compara la coincidència entre els 0 i 1 reals i els 0 i 1 predits per l'algorisme. En Machine Learning, aquest mètode d'avaluació és útil per inspeccionar els errors de cada classe (Beauxis-Aussalet & Hardman, 2014).

El valor de F_1 es defineix com la mitjana harmònica de la precisió i la cobertura (*recall*). La precisió és el valor que comptabilitza el percentatge d'allò que és correcte dins de la predicció (tant 0 com 1). La cobertura comptabilitza la quantitat de positius (1) que l'algorisme ha identificat bé. El propòsit del valor/de la mesura F1 és combinar les mètriques de precisió i cobertura en un sol valor, ja que aquesta mesura ha estat dissenyada per avaluar dades desequilibrades (Korstanje, 2021). Atès que el corpus NewsCom-TOX conté un 89,4% d'instàncies no sarcàstiques i un 10,6% d'instàncies sarcàstiques, s'han avaluat els resultats del SVM amb el valor F1.

Resultats

S'observen a continuació els resultats del SVM entrenat a partir de les dades del corpus NewsCom-TOX. Els valors obtinguts han estat els mateixos tant en l'aplicació del kernel lineal com en el kernel RBF. En la Figura 11, la matriu de confusió mostra que la predicció de l'algorisme ha classificat totes les instàncies noves en la categoria 0 (no sarcàstiques, en aquest cas).

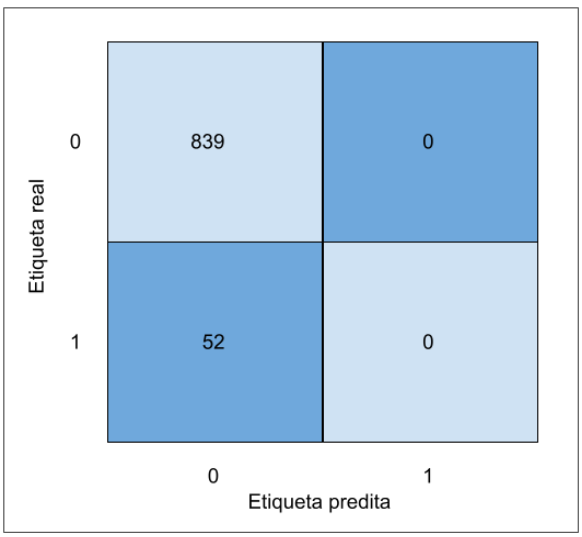


Figura 11. Matriu de confusió. Resultats test data amb SVM (Kernel Linial i Kernel RBF).

.

Aquests resultats molt poc satisfactoris es veuen reflectits en el valor d'F1 a la taula a continuació (Figura 12). S'ha afegit també el valor de la precisió per tal d'il·lustrar la importància d'un mètode adequat d'avaluació que s'adapti a la distribució de les dades del corpus. El valor de la precisió és molt elevat perquè l'algorisme ha predit correctament un 94% de les instàncies de les dades de test (tots "no sarcàstics"). No obstant això, en un conjunt de dades desequilibrat com aquest, el valor de precisió no és fiable, ja que les prediccions correctes no tenen valor si el model no és capaç de classificar correctament cap text en la categoria 1. Per aquesta raó, el percentatge d'èxit del classificador és d'un 0%.

SVM	Precisió (*accuracy*)	F_1
Kernel Linial	0.94	0.00
Kernel RBF	0.94	0.00

Figura 12. Valors de precisió i F_1 en SVM.

Definitivament, els resultats obtinguts no són òptims. El corpus NewsCom-TOX ha estat dissenyat tenint en compte diversos paràmetres de toxicitat, un d'ells el sarcasme, que és un subtipus d'ironia. Malgrat incloure aquest tipus d'ironia, la quantitat d'instàncies sarcàstiques no arriben a ser suficients per a aconseguir bons resultats. A més, segons la definició de sarcasme i ironia establerta per Taulé et al. (2021), és possible que altres expressions irròniques no hagin estat incloses en la categoria "sarcasme" per complir amb els requisits d'aquesta tasca d'anotació concreta. La línia que separa la ironia del sarcasme és molt fina, i la possibilitat que en les dues categories de la data apareguin expressions irròniques podria dificultar la detecció automàtica d'aquestes.

Per tant, davant la dificultat que presenta treballar amb un corpus poc equilibrat i no dissenyat específicament per a l'estudi de la ironia ni per a l'entrenament de classificadors d'ironia en espanyol, s'ha decidit seleccionar unes altres dades més adequades.

4.3 Corpus IRONIA

Per aquesta segona fase del desenvolupament del classificador de la ironia en textos en espanyol, s'han seleccionat i preprocessat 5.000 tuits del corpus desenvolupat per Jasso, i Meza (2016). Aquest corpus en espanyol, conté aproximadament 14.500 tuits irònics i 670.000 tuits no irònics. Per tant, en aquest treball es farà servir un subconjunt del corpus de Jasso i Meza (2016), anomenat com a corpus IRONIA en les pàgines següents.

El mètode d'extracció del corpus es basa en la selecció de tuits escrits en castellà etiquetats com a *#ironia* o *#sarcasmo*, que es consideren irònics sense cap verificació addicional. Aquesta "autoanotació" dels mateixos autors dels tuits no és qüestionada pels autors del corpus, sinó que confien en les etiquetes dels usuaris i assumeixen que són correctes. Ara bé, remarquen que una visió imparcial de l'ús d'aquestes etiquetes a Twitter pot apuntar al que la majoria dels usuaris consideren ironia, i no necessàriament a la definició formal d'aquesta (Jasso i Meza, 2016).

En la seva definició d'ironia, engloben el sarcasme com una subclasse d'aquest fenomen lingüístic. A més, consideren els tuits sarcàstics com a irònics després de verificar manualment que les etiquetes d'*#ironıa* i *#sarcasmo* sovint s'utilitzen indistintament. Quant als tuits no irònics, aquests estan recopilats emprant paraules buides com a termes de cerca (*quién*, *cómo*, *cuándo*, *dónde*, *por qué*, entre altres) evitant alhora els tuits etiquetats com a irònics. És a dir, qualsevol tuit que no està etiquetat explícitament com a irònic és considerat no irònic (Jasso i Meza, 2016). Es mostra en la taula a continuació (Figura 13) un exemple de cada categoria del corpus.

Tuit	Ironia
A veces tengo ganas de irme a vivir a otra provincia o pais, donde no conozca a nadie y nadie me conozca a mi....	0
Calma edurne, aún podemos remontar 🤞👊👌🚀 *#sarcasmo #Edurnevision*	1

Figura 13. Exemples d'instàncies del corpus de Jasso i Meza, (2016).

4.3.1 Obtenció de dades i preprocessament

Recuperació de tuits

En aquest subapartat es descriu el procés de recuperació de tuits del corpus IRONIA. Seguint les limitacions legals de Twitter, el contingut d'aquest corpus es comparteix parcialment. S'ha tingut accés únicament a dues llistes de números d'identificació de cada tuit del corpus, una per als tuits irònics i una per als tuits no irònics (Figura 14). En conseqüència, ha calgut recuperar els tuits a través d'un script en Python (Annex 2).

```
1 605233681779687424
2 605233681377062912
3 605233682274746369
4 605233682262065152
5 605233682551607296
6 605233681762942978
7 605233682648068096
```

Figura 14. Exemples IDs.

Aprofitant un script proposat per Emrah (2019), s'ha fet servir la llibreria Tweepy, que permet accedir fàcilment a l'API (Interfície de Programació d'Aplicacions) de Twitter i extreure els tuits necessaris per a aquesta tasca. Pels requisits temporals del programa, no s'han pogut recuperar la totalitat de tuits del corpus. A més, alguns textos no s'han pogut recuperar a causa de l'antiguitat del corpus. És probable que la plataforma de Twitter o els mateixos usuaris hagin eliminat els tuits en algun moment donat. Així doncs, l'output obtingut és un corpus (Figura 15) que recull aproximadament uns 9.000 tuits irònics i 600.000 tuits no irònics.

```
1 text, ironia
2 No pueden tardar un poquito mas? #sarcasmo 😒👊,1
3 No me salen todas las notificaciones .. Que genial ! #Ironía,1
4 Suena Kurd Maverick como tercer tema para recodar en la #MonsterMusic de tu Monstruo de la Mañana,0
5 @BeliebersFansON Hola! Podrías seguir/recomendar esta página? ------&gt; @Crazy_MofosDms PLOX!💙🤞🏾,0
6 "Tenías razón, ni Cancún ni Playa Del Carmen es tan bonito como los paisajes de las 10 hrs de camino a DF @Tavaresmaxima jaja #Sarcasmo",1
7 "No sé que va a ser de tu vida, nunca haces nada, solo estás acostada - Mi mamá cuando descanso después de que fui su chacha.",0
8 Lindos los reportajes de cierta televisora para la próxima semana sobre el tema familia #sarcasmo,1
```

Figura 15. Arxiu resultant de la recuperació de tuits.

D'entrada, es van seleccionar uns 18.000 tuits (9.000 irònics i 9.000 no irònics) per construir les dades d'entrenament i d'avaluació del classificador. Malauradament, els algorismes implementats amb aquesta quantitat d'instàncies requereixen un temps d'execució molt elevat o una capacitat de CPU, GPU i RAM a la qual no s'ha pogut tenir accés. És per aquesta raó que el conjunt de dades utilitzat finalment és de 5.000 tuits, amb les categories equilibrades (2.500 irònics i 2.500 no irònics) i separats en *train* i *test* amb una proporció 80/20. Aquesta tasca s'ha portat a terme amb el mètode "test_train_split" de la llibreria SkLearn, descrit anteriorment a l'apartat 3.2.1 (Selecció de dades i preprocessament).

Preprocessament: usuaris, hashtags, emoticones i signes de puntuació

Hladka i Holub (2015) insisteixen que les iteracions del cicle de desenvolupament posen a prova les expectatives del desenvolupador i demostren allò que funciona i allò que no, quines funcions excloure i quins paràmetres d'aprenentatge s'estableixen amb quins valors. Malgrat

que no s'ha detallat al llarg de l'apartat de metodologia, la implementació de l'algorisme SVM s'ha provat diverses vegades variant diferents paràmetres, com la mida dels *datatsets (corpus)*, la tokenització i els paràmetres del SVM. En aquest subapartat descrivim el preprocessament del corpus que ha donat millors resultats en l'avaluació final del classificador, realitzat principalment a partir de la llibreria Regex de Python (Annex 3).

En primer lloc, s'han eliminat els usuaris (elements que comencen amb "@") i els enllaços dels documents del corpus, ja que s'ha considerat que no aporten informació útil en la detecció de la ironia. En addició a descartar les etiquetes d'#ironia i #sarcasmo, s'han eliminat també tots els mots que comencen amb "#", perquè és possible que altres etiquetes tinguin relació directa amb la manifestació de la ironia. D'aquesta manera, el classificador no depèn de les etiquetes o "autoanotació" dels usuaris per determinar si el to del tuit és irònic o no.

Tal com s'ha comentat anteriorment, Reyes (1994) proposa que en la ironia, el parlant deforma, exagera o modifica amb burla un enunciat per mostrar una actitud negativa cap aquest o el seu autor. És possible considerar que en el medi escrit, els autors compensen la falta d'informació que l'interlocutor pot inferir a través de la prosòdia o altres elements paralingüístics fent servir certes marques lingüístiques com emoticones o *hashtags* (típiques del medi cibernètic) i signes de puntuació. A més, estudis com els de Kerbrat-Orecchioni (1981) suggereixen la presència de marques lingüístiques com les cometes, els punts suspensius, els signes d'exclamació i l'èmfasi en el to irònic. Així mateix, Reus Boyd-Swan (2009) adverteix la presència de cometes, punts suspensius, signes d'interrogació i exclamació, parèntesis, guions i canvis tipogràfics: cursiva, negreta, versals, etc.

Cal mencionar que, malgrat l'aparició d'aquestes marques lingüístiques no garanteix l'existència del to irònic, és possible que existeixi una coocurrència entre aquestes marques i els enunciats irònics. Encara més, altres estudis que intenten detectar automàticament la ironia també tenen en compte elements estilístics o discursius com els signes de puntuació. Vanin et al. (2013) ja consideraven diversos patrons com alguna expressió estàtica, expressions de riure, puntuació específica i llenguatge simbòlic com emoticones, que possiblement evidencien la presència de missatges irònics. Pinto Cruces (2017) té en compte l'ús d'emoticones, de majúscules, de signes de puntuació i de paraules típiques d'ironia, entre altres.

Per aquestes raons, en aquest treball s'ha mantingut la presència d'emoticones, punts suspensius, interrogants i exclamacions i s'han considerat com a tokens. Per a incorporar les emoticones, s'ha utilitzat la llibreria Emoji de Python, que és capaç de traduir les emoticones del text en la seva transcripció a l'espanyol. Pel que fa als signes de puntuació, han estat transformats amb la llibreria Regex. En una coincidència en el corpus de dos punts seguits ("..") o més, tot el conjunt de punts es tradueix a "*puntos_suspensivos*". En canvi, cadascun dels interrogants i exclamacions s'han considerat com un token individual i s'han traduït a "*signo_interrogación*" i "*signo_exclamación*" respectivament.

D'aquesta manera, la quantitat de tokens d'interrogació i exclamació està vinculada a l'èmfasi que fa servir l'autor del tuit. A més, només s'han traduït a token els signes d'interrogació i exclamació finals ("?" i "!"), ja que en el llenguatge utilitzat en les xarxes socials en espanyol no és comú l'ús de les grafies inicials "¿" i "¡". Així s'evita assignar més pes, o més freqüència, a aquells textos que contenen grafies amb poc valor semàntic. El resultat d'aquest preprocessament del corpus s'exemplifica a continuació (Figura 16).

Tuit original	Tuit preprocessat	Ironia
El autobús tan puntual como siempre.... #ironia	*El autobús tan puntual como siempre* ***puntos_suspensivos***	1
@o2slo @quieee Con 100 cv menos?? Pues corre bastante..!! #sarcasmo #lol	*Con 100 cv menos* ***signo_interrogación signo_interrogación*** *Pues corre bastante* ***puntos_suspensivos signo_exclamación signo_exclamación***	1
y se me va el sueño, cuando entro a twitter 🎶 *http://t.co/WMuW9A6UMo*	*y se me va el sueño, cuando entro a twitter* ***:notas_musicales:***	0
@irstela Ponys que atacan? Nunca habia oido eso... 😂 😂 😂	*Ponys que atacan* ***signo_interrogación*** *Nunca habia oido eso* ***puntos_suspensivos :cara_llorando_de_risa: :cara_llorando_de_risa::cara_llorando_de_risa:***	1

Figura 16. Exemples de tuits preprocessats.

Vectorització de les dades TF/IDF

Per a transformar les dades del corpus IRONIA en la matriu de vectors que entrena el SVM, s'han alterat un parell de paràmetres del mètode TF/IDF Vectorizer per tal d'optimitzar el rendiment del classificador.

- *max_df*: genera un vocabulari de *stopwords* específic del corpus. El vectoritzador ignora els termes que tinguin una freqüència de document estrictament superior al llindar donat, és a dir, al valor de Max_df assignat (Pedregosa et al., 2012). S'ha assignat un Max_df = 0.7 per al processament de les nostres dades.
- *n-gram_range*: determina el nombre de n-grames que s'han d'extreure com a vectors. Després de diverses iteracions amb bigrames i trigrames, s'ha obtingut un millor rendiment del classificador considerant únicament els unigrames.

4.3.2 Support Vector Machine

En aquest apartat es descriu la configuració del SVM i s'avaluen amb una matriu de confusió i el valor F1 els resultats obtinguts a partir de les dades de test del corpus IRONIA (Annex 4).

Implementació model

En la implementació de l'algorisme, també s'han variat alguns paràmetres propis del model. El millor rendiment del classificador ha resultat de la configuració següent:

- Kernel: S'ha fet servir el tipus de kernel no lineal RBF (Radial Basis Function).
- Gamma: és un paràmetre per hiperplans no lineals. Com més elevat el valor de gamma, l'hiperplà s'ajusta més exactament a la separació de les dades d'entreament (Ben Fraj, 2018). S'ha establert el paràmetre de gamma com "*scale*" (per defecte), que calcula el seu valor segons la quantitat de *features* que troba al corpus seguint la formula *1 / (n_features * X.var())* (Pedregosa et al., 2012).
- C: és un paràmetre de penalització d'error. Controla l'equilibri entre un límit de decisió (hiperplà) que sigui homogeni i una classificació acurada de les dades (Ben Fraj, 2018). Per defecte C=1, però amb C=0.1 s'ha trobat un millor rendiment d'aquest classificador de la ironia.

Avaluació model

Es mostren a continuació els resultats de la predicció del SVM entrenat a partir de les dades del corpus IRONIA. La matriu de confusió (Figura 17) reflecteix l'èxit en el rendiment de l'algorisme entrenat amb aquest corpus en comparació amb el corpus NewsCom-TOX. El SVM ha classificat correctament un 91% de les instàncies no irròniques i un 81,8% de les instàncies irròniques. A més, el valor F_1 és 0.8552, notablement alt.

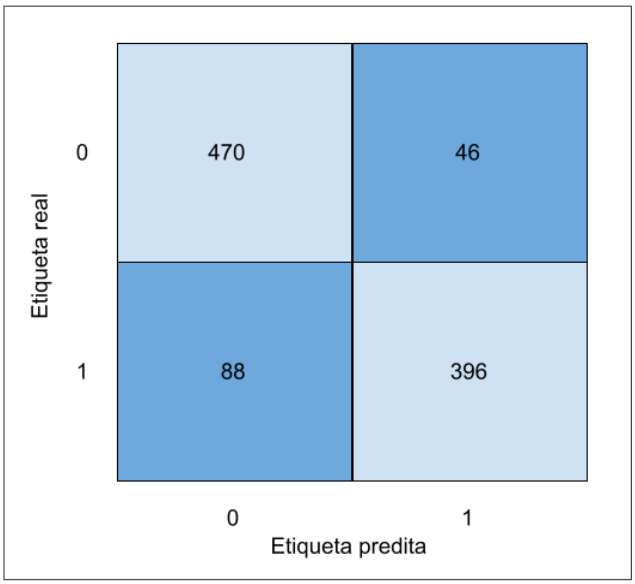


Figura 17. Matriu de confusió. Resultats test data amb SVM (Kernel RBF).

$$\mathbf{F_1 = 0.8552}$$

4.3.3 Xarxa Neuronal: Multilingual Bert

BERT (*Bidirectional Encoder Representations from Transformers*) és un model d'aprenentatge automàtic per al processament del llenguatge natural desenvolupat l'any 2018 per investigadors de Google AI Language. Es pot fer servir per tasques lingüístiques com l'anàlisi de sentiments, cerca de respostes, reconeixement d'entitats nomenades, predicció, generació i resum de textos i desambiguació semàntica, entre altres. L'algorisme BERT ha estat entrenat específicament amb textos de Viquipèdia (uns 2,5 milions de paraules) i de BooksCorpus de Google (uns 800 milions de paraules) (Muller, 2022).

El model BERT fa servir un Transformador, un model de Deep Learning capaç d'aprendre les relacions contextuals entre les paraules d'un text. L'arquitectura bàsica de BERT conté un codificador, que llegeix l'input, i un descodificador que genera les prediccions específiques de

cada tasca. El codificador del Transformador es considera bidireccional perquè llegeix la seqüència de paraules de l'input simultàniament, al contrari que els models direccionals, que el llegeixen de manera seqüencial (d'esquerra a dreta o de dreta a esquerra). Aquesta característica bidireccional permet al model aprendre el context d'una paraula basant-se en els termes que l'envolten (Horev, 2018).

BERT fa servir dues estratègies per trencar les limitacions d'un model direccional: *Masked Language Modeling* (MLM), que aprèn a predir paraules que falten (*masked*) en les oracions segons el context, i *Next Sentence Prediction* (NSP), que rep parells de frases com input i aprèn a predir si la segona frase ha d'aparèixer posteriorment a la primera (Horev, 2018). Per tant, l'input que ha de rebre BERT per ser entrenat és un conjunt d'*embeddings*, o vectors: *token embeddings, segment embeddings,* i *position embeddings* (Muller, 2022) (Figura 18).

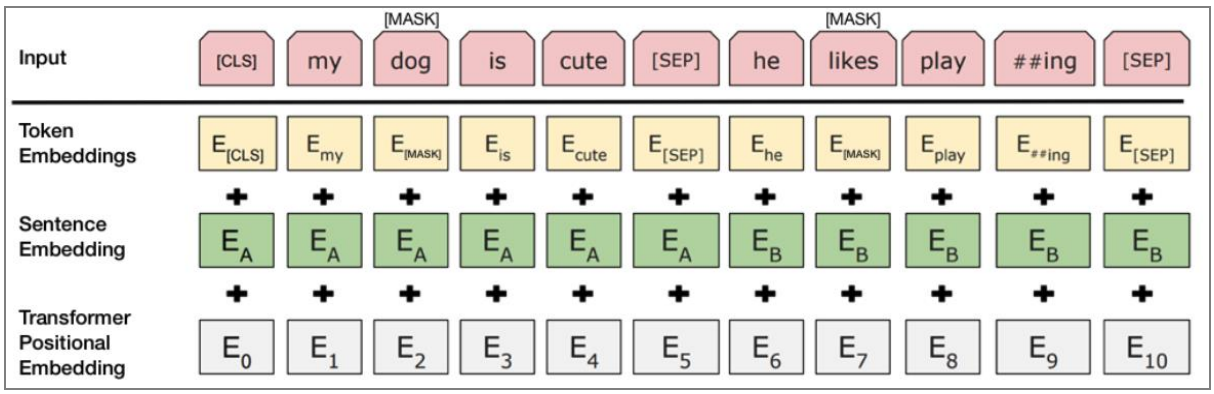


Figura 18. Representació d'input BERT (Horev, 2018).

Implementació model

Per aquest treball, s'ha explorat el desenvolupament d'un classificador de la ironia en tuits en espanyol amb la implementació del model *Multilingual BERT (cased)*. Aquest model ha estat preentrenat igual que BERT original, i utilitza un vocabulari extret de la Viquipèdia multilingüe. A més, el terme *cased* fa referència al fet que en aquest model les entrades de text conserven la distinció entre majúscules i minúscules i els accents (*TensorFlow Hub,* s.d.).

Per començar, s'han vectoritzat les dades ja preprocessades, tal com s'explica en el subapartat 4.3.1 i s'ha construït el model d'aprenentatge automàtic fent servir el transformador Multilingual Bert a través de les llibreries de TensorFlow i Keras (Annex 5).

La xarxa neuronal construïda conté dues capes, i s'ha configurat el model perquè entreni amb les dades en un total de 10 iteracions, també anomenades *epochs*. La quantitat d'*epochs* determina les vegades que l'algorisme entrenarà amb tot el conjunt de dades d'entrenament. Després de cada iteració, s'actualitzen els paràmetres interns del model. Aquest procés permet al model aprendre i millorar el seu rendiment amb un mateix grup de dades (Brownlee, 2018). Malgrat que el nombre d'*epochs* pot ser elevat (centenars o milers) per a diferents tasques, en la realització d'aquest treball únicament s'han configurat 10, a causa de les limitacions temporals i de recursos disponibles.

Avaluació model

Per a l'avaluació del model, s'han fet servir altres llibreries com Numpy, Sklearn i matplotlib. Els resultats aconseguits a partir de Multilingual BERT es representen a continuació amb una matriu de confusió (Figura 19), on s'observa que aquest model tendeix a classificar aproximadament la mateixa quantitat d'instàncies correctament en cada categoria. El valor F1 ha resultat en 0.5943, ja que un gran nombre d'instàncies han estat classificades erròniament.

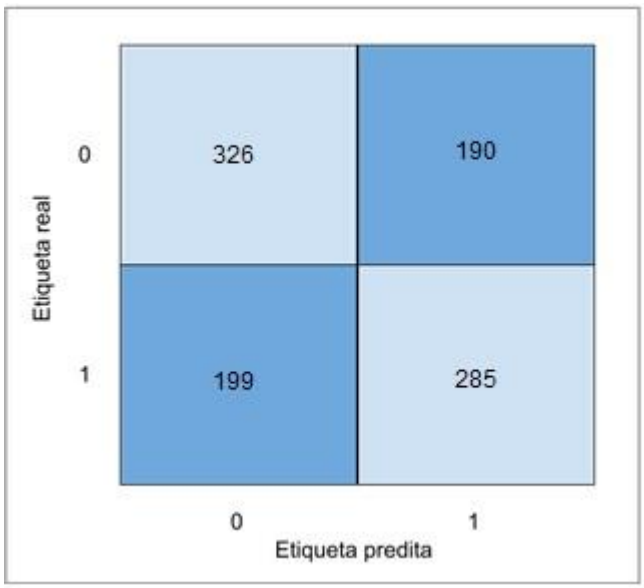


Figura 19. Matriu de confusió. Resultats test data de BERT multilingüe.

F_1 = 0.5943

5. ANÀLISI DELS RESULTATS

En aquest apartat es comparen els valors F_1 dels dos models d'aprenentatge automàtic entrenats a partir del corpus IRONIA en tuits en espanyol, es proposen altres mètodes o tècniques per la resolució d'aquesta tasca i es comparen els resultats obtinguts en aquest treball amb altres estudis recents.

Model	F_1
SVM (kernel RBF)	**0.8552**
Multilingual BERT (cased)	**0.5943**

Figura 20. Comparació de valors F_1.

El SVM ha estat l'algorisme amb millors resultats (Figura 20), amb un valor F_1 de 0,85. Malgrat aquests bons resultats, val la pena considerar les limitacions d'algunes de les tècniques emprades durant el procés d'aprenentatge automàtic. S'ha fet servir el Vectoritzador TF/IDF, que és útil únicament a nivell lèxic i no considera la semàntica (per tant, ignora els sinònims), ni les seqüències de les paraules (Ramadhan, 2021).

Pel que fa a la semàntica, diversos estudiosos de l'anàlisi del discurs en l'espanyol indiquen la presència de figures retòriques en la manifestació de la ironia. Muecke (1970) assenyala la tendència a extremar hiperbòlicament l'èmfasi de les afirmacions durant el to irònic, de manera que el receptor adverteixi l'excés i entengui la dissonància entre les dues veus del discurs irònic o sarcàstic. Així mateix, Kerbrat-Orecchioni (1981) ressalta la presència de figures retòriques en la manifestació de la ironia en els textos escrits. Assenyala la hipèrbole, l'oxímoron, la lítote, la metàfora i la paradoxa, a més de remarcar la possible existència de contradicció entre dos segments de l'enunciat, la contradicció d'un fet evident per als interlocutors. Finalment, Schoentjes (2003) esmenta marques de caràcter lingüístic com la hipèrbole, l'oxímoron, la lítote, la metàfora i la paradoxa. La taula a continuació recull les figures retòriques trobades a les dades analitzades, la hipèrbole i la metàfora.

A causa de les limitacions temporals d'aquest treball, no s'ha pogut incorporar una aproximació semàntica al preprocessament en les dades emprades per entrenar els models. Seria interessant considerar en el futur l'addició de nous *features* utilitzant mètodes com el *POS tagging* o l'anàlisi de sentiments, que podrien millorar el rendiment del SVM o de BERT.

Pel que fa al Multilingual BERT, és probable que el valor d'F1 sigui baix perquè el model està dissenyat per aprendre millor amb grans quantitats de dades i un gran nombre d'*epochs*. Segons Horev (2018), els resultats de BERT milloren conforme la quantitat de dades d'entrenament augmenta. Per tant, seria interessant entrenar aquest model amb un conjunt de dades més gran i amb una afinació dels paràmetres (*fine-tuning*) més exhaustiva. A més, es podrien obtenir també resultats interessants a partir de la implementació d'altres models com BETO (Cañete et al., 2020), un model BERT entrenat exclusivament en textos en espanyol; DistilBERT, la versió més ràpida de BERT entrenada amb gairebé la meitat de paràmetres que l'original; o fins i tot un Random Forest o un model de Regressió Logística.

Es reflexiona també sobre l'antiguitat del corpus emprat. Els models descrits en aquest treball han estat entrenats amb dades recollides l'any 2016. Considerant la ràpida evolució del llenguatge emprat en el medi cibernètic, és possible que les estratègies discursives dels usuaris de les xarxes canviïn constantment. Això podria implicar una necessitat d'actualització de les dades d'entrenament cada poc temps.

Finalment, es mostra en la taula següent (Figura 21) una comparació entre els diferents resultats publicats en Alnajjar i Hämäläinen (2021) i Ortega-Bueno et al. (tasca IrosVA, 2019), Jasso i Meza (2016) i els resultats obtinguts en aquest treball. Alnajjar i Hämäläinen (2021) han obtingut resultats excel·lents, tant amb la implementació d'un SVM i un model neuronal, a partir de transcripcions de capítols de sèries de televisió (2 capítols de *South Park* i 2 capítols d'*Archer*) en espanyol peninsular i espanyol llatinoamericà. Jasso i Meza (2016) obtenen també bons resultats amb l'implementació d'un SVM i 32.000 tuits. Ortega-Bueno et al. (2019) descriuen els resultats dels grups guanyadors de la tasca IroSvA (ELiRF-UPV i CIMAT), que implementen transformadors per abordar la tasca de detecció automàtica de la ironia en dialectes peninsular, mexicà i cubà. Les diferències que existeixen entre la tipologia i la mida del corpus, l'ús de diferents variants de la llengua espanyola, els diferents models implementats i els diferents valors d'avaluació, que varien entre F_1 i Precisió, dificulten la comparació entre resultats. No obstant això, es pot observar que els resultats obtinguts a partir del SVM en aquest treball s'apropen molt als resultats obtinguts per Jasso i Meza (2016), estudi del qual s'ha extret un fragment de les dades per construir els classificadors descrits en els apartats anteriors. En canvi, respecte dels models neuronals, els resultats de la tasca IroSvA superen els resultats d'aquest treball.

Corpus/autors	Font	Dades	Varietat de l'espanyol	SVM	Xarxa Neuronal
Alnajjar i Hämäläinen (2021)	Transcripció sèries TV	4 capítols	Espanyol peninsular i espanyol llatinoamericà	89.0% (precisió)	87,5% (precisió)
Jasso i Meza (2016)	Twitter	32.000 tuits	Dialectes diversos	86.0% (F_1)	-
Corpus IRONIA (Jasso i Meza, 2016)	Twitter	5.000 tuits	Dialectes diversos	**85.52% (F_1)**	**59,43% (F_1)**
ELiRF-UPV (Ortega-Bueno et al., IrosVA, 2019)	Twitter	3.000 tuits	Espanyol peninsular	-	71.67% (F_1)
ELiRF-UPV (Ortega-Bueno et al., IrosVA, 2019)	Twitter	3.000 tuits	Espanyol mexicà	-	68.03% (F_1)
CIMAT (Ortega-Bueno et al., IrosVA, 2019)	Twitter	3.000 tuits	Espanyol cubà	-	65.96% (F_1)

Figura 21. Comparació de resultats entre diversos estudis

6. CONCLUSIONS

L'objectiu general d'aquest treball ha estat aprendre com és el procés de desenvolupament d'un classificador que resol una tasca lingüística aplicant tècniques d'aprenentatge automàtic i eines de PLN. Aquest treball ha permès experimentar de primera mà la recuperació de dades i gestió del corpus, el processament de les dades, la implementació i l'avaluació dels models SVM i Multilingual BERT.

No solament s'han pogut ampliar els coneixements sobre el llenguatge de programació Python i diferents eines de PLN, sinó que s'han aconseguit resoldre problemes i dificultats que han aparegut al llarg de tot el procés d'aprenentatge automàtic. S'ha comprovat la importància del disseny, adequació i processament dels corpus i l'impacte d'aquests en el

resultat final de l'algorisme. Finalment, s'ha reconegut la necessitat d'un bon domini de les tècniques de PLN i aprenentatge automàtic a més d'una comprensió teòrica del fenomen lingüístic a estudiar.

Pel que fa a l'objectiu concret d'aquest treball, s'han implementat dos models per a la classificació de la ironia en textos en espanyol, que apliquen tècniques d'aprenentatge automàtic supervisat diferents. Per una banda, el model SVM, amb una aproximació geomètrica a les tasques d'aprenentatge automàtic, amb el qual s'ha obtingut un valor F_1 de 0.8552. Per l'altra banda, s'ha obtingut un F_1 de 0.5943 amb el transformador BERT multilingüe, un model de Deep Learning capaç d'aprendre les relacions contextuals entre les paraules d'un text. Aquests resultats encoratjadors confirmen la hipòtesi de partida: és possible classificar automàticament els textos en espanyol segons si contenen una ironia o no, aplicant mètodes d'aprenentatge automàtic supervisat. A més, els resultats obtinguts durant les diferents fases del desenvolupament d'aquest treball posen en relleu la necessitat de disposar de corpus amb dades equilibrades i dissenyats i anotats específicament per la tasca formulada.

Per acabar, és important remarcar la dificultat que presenta la detecció automàtica de la ironia o el sarcasme, i en general del llenguatge figurat. Segons Reyes, "la ironia és una reflexió, més o menys complexa, sobre la realitat, sobre la relació entre el llenguatge i la realitat i sobre la relació entre una frase i els usos previs d'aquesta frase". Per tant, "la interpretació de la ironia requereix coneixements sobre el món, sobre el parlant i sobre la relació entre el parlant i l'oient". A més, Warning (citat per Reus Boyd-Swan, 2009) afirma que els senyals irònics pertanyen al nivell de la *parole* i no de la *langue*, i deuen la seva identificació a un coneixement previ de caràcter pragmàtic: un conjunt de pressuposicions que el receptor ha de tenir presents perquè s'interpreti adequadament el missatge. Aquestes reflexions lingüístiques inviten a considerar la tasca de la detecció automàtica de la ironia com un problema que involucra la pragmàtica, a més de les característiques estructurals dels textos. Indubtablement, un dels reptes actuals de la lingüística computacional i la PLN és intentar trobar un mètode efectiu per la detecció automàtica de la ironia, malgrat les dificultats lingüístiques i tecnològiques que presenta aquesta tasca.

7. BIBLIOGRAFIA

Alnajjar, K., i Hämäläinen, M. (2021). ¡*Qué maravilla! Multimodal Sarcasm Detection in Spanish: a Dataset and a Baseline*. http://arxiv.org/abs/2105.05542

Barbieri, E., Barbieri, F., Ronzano, F., i Saggion, H. (2015). *Procesamiento del Lenguaje Natural Sociedad Española para el Procesamiento del Lenguaje Natural Is This Tweet Satirical? A Computational Approach for Satire Detection in Spanish* *. 135-142. http://www.redalyc.org/articulo.oa?id=515751524015

Beauxis-Aussalet, E., i Hardman, L. (2014). *Visualization of Confusion Matrix for Non-Expert Users*.

Ben Fraj, M. (2018). *In Depth: Parameter tuning for SVC*. Recuperat 20 juliol 2022, de https://medium.com/all-things-ai/in-depth-parameter-tuning-for-svc-758215394769

Brownlee, J. (2018). *Difference Between a Batch and an Epoch in a Neural Network*. Recuperat 20 juliol 2022, de https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/

Brownlee, J. (2020). *Train-Test Split for Evaluating Machine Learning Algorithms*. Recuperat 20 juliol 2022, de https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

Calvo, H., Gambino, O. J., i Mendoza, C. V. G. (2020). Irony detection using emotion cues. *Computacion y Sistemas*, *24*(3), 1281-1287. https://doi.org/10.13053/CYS-24-3-3487

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., i Pérez, J. (2020). *Spanish Pre-Trained BERT Model and Evaluation Data*. Recuperat 15 juliol 2022, de https://github.com/josecannete/spanish-corpora

Carvalho, P., Sarmento, L., Silva, M. J., i Oliveira, E. (2009). Clues for detecting irony in user-generated contents: Oh...!! it's «so easy» ;-). *International Conference on*

Information and Knowledge Management, Proceedings, 53-56. https://doi.org/10.1145/1651461.1651471

Cignarella, A. T., Basile, V., Sanguinetti, M., Bosco, C., Rosso, P., i Benamara, F. (2020). *Multilingual Irony Detection with Dependency Syntax and Neural Models*. http://arxiv.org/abs/2011.05706

Crystal, D. (2011). *Internet Linguistics*. Routledge. https://doi.org/10.4324/9780203830901

DETOXIS-IberLEF. (2021). *DETOXIS-IberLEF 2021*. Recuperat 15 juliol 2022, de https://detoxisiberlef.wixsite.com/website

Emrah, A. (2019). *Fetch Tweets Using Their IDs With Tweepy, Twitter API and Python*. Recuperat 10 maig 2022, de https://medium.com/analytics-vidhya/fetch-tweets-using-their-ids-with-tweepy-twitter-api-and-python-ee7a22dcb845

Farías, D. I. H., Patti, V., i Rosso, P. (2016). Irony Detection in Twitter. *ACM Transactions on Internet Technology*, *16*(3), 1-24. https://doi.org/10.1145/2930663

González-Ibáñez, R., Muresan, S., i Wacholder, N. (2011). Identifying Sarcasm in Twitter: A Closer Look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 581-586. http://www.vidarholen.net/contents/interjections/

Herring, S. C. (Ed.). (1996). *Computer-Mediated Communication* (Vol. 39). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.39

Hladka, B., i Holub, M. (2015). A gentle introduction to machine learning for natural language processing: How to start in 16 practical steps. *Language and Linguistics Compass*, *9*(2), 55-76. https://doi.org/10.1111/lnc3.12123

Horev, R. (2018). *BERT Explained: State of the art language model for NLP*. Recuperat 15 juliol 2022, de https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Irazú, D., Irazú, I., Farías, H., Sulis, E., Patti, V., Ruffo, G., i Bosco, C. (2015). *ValenTo: Sentiment Analysis of Figurative Language Tweets with Irony and Sarcasm *.* http://www.cs.uic.edu/

Jasso, G., & Meza, I. (2016). *Character and Word Baselines for Irony Detection in Spanish Short Texts *.*

Kerbrat-Orecchioni, C. (1981). L'ironie comme trope. *Poétique*, 108-127.

Korstanje, J. (2021). *The F1 score: Towards Data Science*. Recuperat 15 juliol 2022, de https://towardsdatascience.com/the-f1-score-bec2bbc38aa6

Kunneman, F., Liebrecht, C., van Mulken, M., i Van den Bosch, A. (2015). Signaling sarcasm: From hyperbole to hashtag. *Information Processing & Management*, *51*(4), 500-509. https://doi.org/10.1016/j.ipm.2014.07.006

Muecke, D. (1970). *Irony*. Methuen.

Muller, B. (2022). *BERT 101 - State Of The Art NLP Model Explained*. Recuperat 15 juliol 2022, de https://huggingface.co/blog/bert-101#2-how-does-bert-work

Navlani, A. (2019). *Scikit-learn SVM Tutorial with Python (Support Vector Machines). DataCamp*. Recuperat 17 juliol 2022, de https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python

Ortega-Bueno, R., Rangel, F., Irazú, D., Farías, H., Rosso, P., Montes-Y-Gómez, M., i Medina-Pagola, J. E. (2019). *Overview of the Task on Irony Detection in Spanish Variants*. https://pakdd16.wordpress.fos.auckland.ac.nz/technical-program/contests/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V.,

Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., i Duchesnay, É. (2012). *Scikit-learn: Machine Learning in Python*.

Potamias, R. A., Siolas, G., i Stafylopatis, A. G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications, 32*(23), 17309-17320. https://doi.org/10.1007/s00521-020-05102-3

Ramadhan, L. (2021). *TF-IDF Simplified. A short introduction to TF-IDF: Towards Data Science*. Recuperat 15 juliol 2022, de https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530

Reus Boyd-Swan, F. (2009). Cómo se manifiesta la ironía en un texto escrito. A P. Lang (Ed.), *Dime cómo ironizas y te diré quién eres: una aproximación pragmática a la ironía* (Vol. 45, p. 293-308).

Reyes, G. (1994). *Los Procedimientos de cita: citas encubiertas y ecos*. Arco/Libros.

Ruiz Gurillo, L. (2012). *La Lingüística del humor en español: cómo convencer con palabras*. Arco/Libros.

Schoentjes, P. (2003). *La poética de la ironía*. Cátedra.

Taulé, M., Ariza, A., Nofre, M., Amigó, E., i Rosso, P. (2021). Overview of DETOXIS at IberLEF 2021: DEtection of TOXicity in comments In Spanish. En *Procesamiento de Lenguaje Natural* (Vol. 67, p. 209-221). Sociedad Espanola para el Procesamiento del Lenguaje Natural. https://doi.org/10.26342/2021-67-18

TensorFlow Hub. (s.d.). Recuperat 20 juliol 2022, de https://tfhub.dev/tensorflow/bert_multi_cased_preprocess/3

Vanin, A. A., de Freitas, L. A., Vieira, R., i Bochernitsan, M. N. (2013). *Some Clues on Irony Detection in Tweets*. http://twitter4j.org/en/index.html

Zhang, C., i Abdul-Mageed, M. (2019). *Multi-Task Bidirectional Transformer Representations for Irony Detection*. http://arxiv.org/abs/1909.03526